国際標準化糖鎖線形表記法～国際糖鎖構造データリポジトリへ向けて～

○山田一作[1]、木下聖子[2,3]、松原正陽[1]、土屋伸一郎[2]、新町大輔[2]、藤田晶大[2]、青木信幸[2]、鹿内俊秀[3]、鈴木芳典[3]、ソロビヨワ・イェレナ[3]、藤田典昭[3]、奥田修二郎[4]、川嵜敏祐[5]、成松久[3]
[1]野口研究所、[2]創価大学、[3]産業技術総合研究所・糖鎖創薬技術研究センター、[4]新潟大学、[5]立命館大学

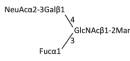
トーゴーの日シンポジウム2014～第2段階に入ったライフサイエンスデータベースの統合～」@Tokyo

Abstract

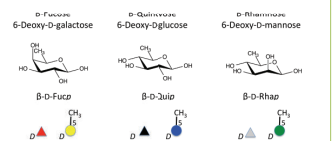
近年、World Wide Webの利便性を向上させ得るセマンティックウェブはライフサイエンスデータを連携するための手段として注目されている。糖鎖研究により生産された各種データに関しても、この新潮流に対応した開発が進んでいる。この流れを受けて糖鎖構造をResource Description Framework(RDF)をベースとしたセマンティックウェブに対応するための新たな線形表記法としてWURCS（Web3 Unique Representation of Carbohydrate Structures）を開発した。WURCSの特徴は、1）URIとして可能な線形文字列表記、2）グリコシド結合が不明確など曖昧さを含む糖鎖構造を一意に文字列表記　である。WURCSは、アグリコンを除いた糖鎖構造を線形文字列として記述するため、各種データベースに含まれる分子をWURCSへ変換することにより、共通の糖鎖構造を含む分子を検索することができる。また、国際糖鎖構造データリポジトリ開発におけるWURCS活用について紹介する。

Sequence Notation

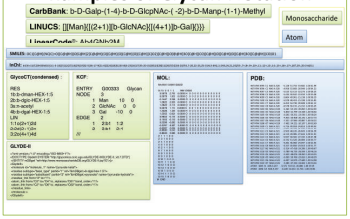
- **Nucleic acid**
  TGATTAGTGATTAGTGATTAGTGATTA
- **Protein, Peptide**
  MKKISRKEYVSMYGPTTGDKVRLGDTDLIAEVEH
- **Glycan**
  NeuAcα2-3Galβ1-4(Fucα1-3)GlcNAcβ1-2Manα1-R



Notation of Monosaccharide



Examples of Glycan Notation



WURCS

Web3 Unique Representation of Carbohydrate Structures

linear notation
unique notation
- nonstandard monosaccharide units
- compositions
- repeating units
- ambiguous structures
- linkages/linkage positions

J. Chem. Inf. Model. 2014, 54, 1558−1566

Definition-1

Backbone ... the carbon backbone of a monosaccharide
Aglycone ...a nonsugar component of a glycan usually located at the C-1 anomeric position
Modification ...a nonsugar component that modifies a sugar component
Glycan ... a component of a molecule consisting of backbones and their modifications
NOS* ... an N, O or S atom
Version ... The version number of WURCS.

Definition-2

Among the three parts extracted from the **glycan**, the **backbones**, **modifications** and attachment information between the backbones and modifications are represented as strings called SkeletonCodes, ALINs and COLINs, respectively. Each of these is formally defined as follows.

SkeletonCode and **SkeletonCode Character**
- SkeletonCode is a string representation of backbone carbons including the adjacent atoms in modifications. Each character of a SkeletonCode is called a SkeletonCode character.

ALIN, Atomic Linear Notation
COLIN, COnnection Linear Notation
- A string representing the connection between a backbone and modification. A COLIN is composed of the followingfour units:
- **NB**: Numbering of Backbones
- **PCB**: Position of the connected Carbon t the Backbone.
- **DMB**: Direction of Modification on the Backbone carbon.
- **PCA**: Position of connected backbone Carbon in ALIN.

- Using these four units, a COLIN is represented as follows:
  - **NB+PC3:DMB-PCA**
- A COLIN is listed just before the corresponding ALIN. The COLINs may be compressed, as described later.

Definition-3

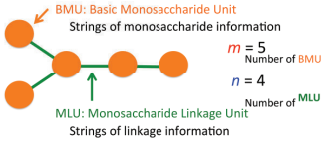
MOD, linear notation of MODification
- It is composed of COLINs and ALIN. If multiple COLINs compose the modification, they are listed consecutively using "," as a delimiter. There is no delimiter between COLIN and ALIN.A MOD is thus represented follows:
  - **COLIN1,..,COLINkALIN**
- where k indicates the number of COLINs.

Version, the version number of WURCS
BMU, Basic Monosaccharide Unit
- MODs are listed after the SkeletonCode, separated by a vertical bar "I". A BMU is thus represented follows:
  - **SkeletonCodeIMOD1I...IMODI**
- where I indicates the number of MODs.For example, GlcNAc would be represented as follows:
  - **12122h1+1:1-1,1+52-2*O*1!+2:2-1*NCC/3=O**

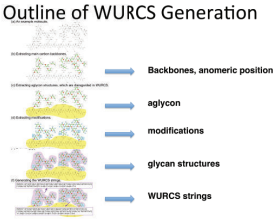
MLU, Monosaccharide Linkage Unit

WURCS Format

WURCS=*Version*/*m*,*n*/[BMU1]...[BMU*m*]MLU1|...|MLU*n*

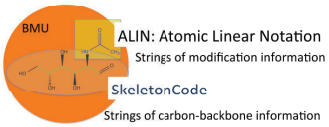


Outline of WURCS Generation

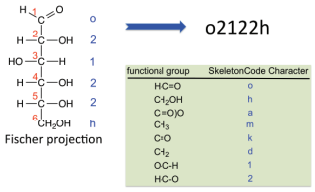


BMU: Basic Monosaccharide Unit

Strings of Monosaccharide

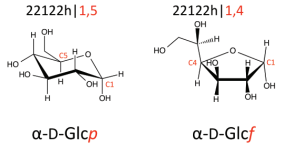


SkeletonCode

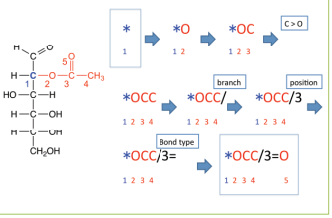


functional group	SkeletonCode Character
HC=O	o
C_{1}OH	h
C=O)O	a
C_{1}3	m
C>O	k
C_{1}2	d
OC-H	1
HC-O	2

Representation of Ring Form



ALIN Generation of Acetyl Group



Symbols of ALIN

Atom	Symbol
Backbone Carbon	*
Carbon	C
Nitrogen	N
Oxygen	O
Hydrogen	
...	...

	Symbol
Position	1, 2, 3, ...
Direction	U, L, /, \, -
Stereo	^E, ^Z, ^R, ^S

Type	Symbol	example
Branch	/	*OCC/3=O
Connection	$	*CCCCCCS2

Bond	Symbol	example
Single		
Double	=	
Triple	#	
Aromatic	(,)	*(CCCCCCS2)

ALIN: Atomic Linear Notation

abbreviation	substituent group	ALIN
Ac	acetyl	*OCC/3=O
NAc	*N*-acetyl	*NCC/3=O
P	phosphate	*OPO/3O/3=O
S	sulfate	*OSO/3=O/3=O
N	Primary amine	*N
NGc	*N*-glycolyl	*NCCO/3=O

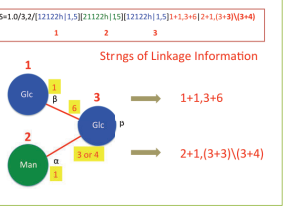
WURCS of Monosaccharides

Monosaccharide	WURCS
aldehydo-D-Glc	WURCS=1.0/1,0,0/a2122h
α-D-Glcp	WURCS=1.0/1,0,0/a2122h-1,5]
β-D-Glcp	WURCS=1.0/1,0,0/a2122h-1,5]
β-D-Glcf	WURCS=1.0/1,0,0/a2122h-1,4]
α-D-Manp	WURCS=1.0/1,0,0/a2112h-1,5]
β-D-Manf	WURCS=1.0/1,0,0/a1122h-1,4]
β-D-Galp	WURCS=1.0/1,0,0/a2112h-1,5]
α-L-Fucp	WURCS=1.0/1,0,0/a1221m-1,5]
β-D-GlcpN	WURCS=1.0/1,0,0/a2122h-1,5,1*N]
α-D-NeupSAc	WURCS=1.0/1,0,0/a2d2112h-[2,6]2*O[5*NCC/3=O]
β-D-GlcpNAc	WURCS=1.0/1,0,0/a2122h-1,5]2*NCC/3=O]
β-D-GalpNAc	WURCS=1.0/1,0,0/a2112h-1,5]2*NCC/3=O]
β-D-GalpA	WURCS=1.0/1,0,0/a2112a-1,5]
L-IdopA	WURCS=1.0/1,0,0/a2121a-1,5]

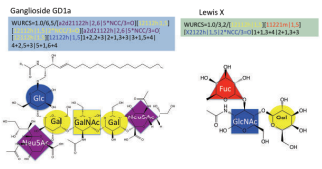
MLU: Monosaccharide Linkage Unit

WURCS=1.0/3,2/[12122h|1,5][21122h|1,5][12122h|1,5]1+1,3+6|2+1,(3+3)\(3+4)

Strngs of Linkage Information



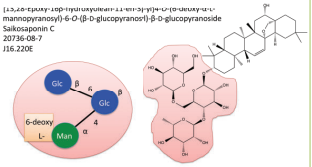
WURCS of GD1a, Lewis X



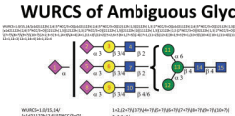
Crude Drug : Saikosaponin C

WURCS=1.0/3,2/[12122h|1,5][1221m|1,5][12122h|1,5]1+1,3+6|2+1,3+4

- Saikosaponin C
- 20736-08-7
- J16.220E



WURCS of Ambiguous Glycan



Number of Structures in Various DB Converted to WURCS

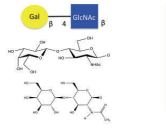
Database	Number of compounds	Distinct number or WURCS	Number of multiple WURCS
ChEBI	4827	4648	218
PDB	3192	722	0
KEGG Compound	2429	2322	168
GlycoNAVI	3492	3453	93
JMSDB	886	839	0
LFDB	186	186	0

Distinct number of WURCS reflects the number of unique glycan structures in the database. Number of multiple WURCS reflects the number of database entries containing more than one unique glycan structure.

DB Search by WURCS String

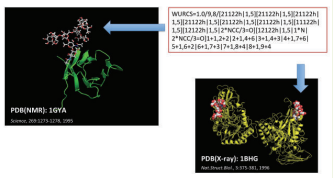
LacNAc

WURCS=1.0/2,1,1/[12112h|1,5|12122h|1,5|2*NCC/3=O]1+1,2+4



ChEBI	ChEBI:16153
ChEBI	ChEBI:62124
KEGG	C00611
KEGG	C00694
KEGG	C04657
KEGG	C04791
KEGG	C04818
PDB	1a3k
PDB	1kjl
PDB	2d6n
PDB	2wt0

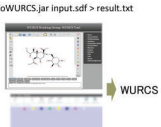
PDB Entries with Same Glycan Structure



WURCS Tools

Command Line Tool
$ java -jar -sdf ./MolfileToWURCS.jar input.sdf > result.txt

Web Application
- Molfile
- Chemical structure
- Glycan structure



WURCS Working Group

- http://www.wurcs-wg.org

Acknowledgements

This research was supported by Japan Science and Technology Agency (JST), National Bioscience Database Center (NBDC) and National Institute of Advanced Industrial Science and Technology (AIST) in Japan.

Licensed under a Creative Commons 表示2.1日本 license (c)2014 山田一作（公益財団法人野口研究所）